Inferring properties of disordered chains from FRET transfer efficiencies

Zheng, Wenwei ; Zerze, Gül H ; Borgia, Alessandro ; Mittal, Jeetain ; Schuler, Benjamin ; Best, Robert B

Abstract: Förster resonance energy transfer (FRET) is a powerful tool for elucidating both structural and dynamic properties of unfolded or disordered biomolecules, especially in single-molecule experiments. However, the key observables, namely, the mean transfer efficiency and fluorescence lifetimes of the donor and acceptor chromophores, are averaged over a broad distribution of donor-acceptor distances. The inferred average properties of the ensemble therefore depend on the form of the model distribution chosen to describe the distance, as has been widely recognized. In addition, while the distribution for one type of polymer model may be appropriate for a chain under a given set of physico-chemical conditions, it may not be suitable for the same chain in a different environment so that even an apparently consistent application of the same model over all conditions may distort the apparent changes in chain dimensions with variation of temperature or solution composition. Here, we present an alternative and straightforward approach to determining ensemble properties from FRET data, in which the polymer scaling exponent is allowed to vary with solution conditions. In its simplest form, it requires either the mean FRET efficiency or fluorescence lifetime information. In order to test the accuracy of the method, we have utilized both synthetic FRET data from implicit and explicit solvent simulations for 30 different protein sequences, and experimental single-molecule FRET data for an intrinsically disordered and a denatured protein. In all cases, we find that the inferred radii of gyration are within 10% of the true values, thus providing higher accuracy than simpler polymer models. In addition, the scaling exponents obtained by our procedure are in good agreement with those determined directly from the molecular ensemble. Our approach can in principle be generalized to treating other ensemble-averaged functions of intramolecular distances from experimental data.

DOI: https://doi.org/10.1063/1.5006954

Originally published at:
Zheng, Wenwei; Zerze, Gül H; Borgia, Alessandro; Mittal, Jeetain; Schuler, Benjamin; Best, Robert B (2018). Inferring properties of disordered chains from FRET transfer efficiencies. Journal of Chemical Physics, 148(12):123329.
DOI: https://doi.org/10.1063/1.5006954

Inferring Properties of Disordered Chains from FRET Transfer Efficiencies

Wenwei Zheng[1,a,b], Gul H. Zerze[2], Alessandro Borgia[3], Jeetain Mittal[2], Benjamin Schuler[3,4,a], Robert B. Best[1,a]

[1]*Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520, USA.*

[2]*Department of Chemical and Biomolecular Engineering, Bethlethem, PA 18015, USA*

[3]*Department of Biochemistry and* [4]*Department of Physics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland*

Förster resonance energy transfer (FRET) is a powerful tool for elucidating both structural and dynamic properties of unfolded or disordered biomolecules, especially in single-molecule experiments. However, the key observables, namely the mean transfer efficiency and fluorescence lifetimes of the donor and acceptor chromophores, are averaged over a broad distribution of donor-acceptor distances. The inferred average properties of the ensemble therefore depend on the form of the model distribution chosen to describe the distance, as has been widely recognized. In addition, while the distribution for one type of polymer model may be appropriate for a chain under a given set of physico-chemical conditions, it may not be suitable for the same chain in a different environment, so that even an apparently consistent application of the same model over all conditions may distort the apparent changes in chain dimensions with variation of temperature or solution composition. Here, we present an alternative and straightforward approach to determining ensemble properties from FRET data, in which the polymer scaling exponent is allowed to vary with solution conditions. In its simplest form, it requires either the mean FRET efficiency or fluorescence lifetime information. In order to test the accuracy of the method, we have utilized both synthetic FRET data from implicit and explicit solvent simulations for 30 different protein sequences, and experimental single-molecule FRET data for an intrinsically disordered and a denatured protein. In all cases, we find that the inferred radii of gyration are within 10 % of the true values, thus providing higher accuracy than simpler polymer models. In addition, the scaling exponents obtained by our procedure are in good agreement with those determined directly from the molecular ensemble. Our approach can in principle be generalized to treating other ensemble-averaged functions of intramolecular distances from experimental data.

I. INTRODUCTION

It is increasingly realized that disorder plays a key role in biology, exemplified by the rich variety of functions performed by intrinsically disordered proteins, from transcription regulators to molecular chaperones [1]. Studying the structure, dynamics and function of these disordered polypeptide chains, as well as other disordered biopolymers, requires new experimental methods, and new methods of interpreting the data, since they are inherently averaged over a broad ensemble of heteropolymer configurations[2-4]. Examples of experiments which can help address this challenge include: nuclear magnetic resonance (NMR), which can provide short-range structural information via scalar coupling, chemical shift and NOE data[5], as well as long-range

[a)] Electronic mail: wenweizheng@asu.edu, schuler@bioc.uzh.ch, robertbe@helix.nih.gov.

[b)] Current address: College of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona 85212, USA.

information by paramagnetic relaxation enhancement (PRE) measurements[6], and global information from diffusion coefficient measurements; light-scattering[7] and two-focus FCS[4, 8], which also yield diffusion coefficients and hence hydrodynamic radius; small-angle X-ray (or neutron) scattering (SAXS or SANS), which directly gives information on inter- and intramolecular pair distance distributions [9]; and Förster resonance energy transfer (FRET), in particular single-molecule FRET, which probes the distribution of distances between pairs of chromophore labels attached to the molecule of interest, as well as the associated dynamics, and often enables structured and unstructured subpopulations to be separated [10, 11]. Obtaining as detailed as possible a picture of the disordered ensemble would ideally combine information from all of these experiments, if available. This can be done most comprehensively via an explicit ensemble description of the disordered state, either by reweighting of an existing molecular simulation[2-4, 12-17], or by performing an ensemble structural refinement with a very large number of replicas of the system[18-23]. However, a simpler approach is frequently useful. For example, one may have limited experimental data available so that the result from ensemble simulation or reweighting methods would be strongly dependent on the quality of the simulation and force field used. In this case, a more straightforward analysis of the data, which does not involve running extensive simulations, may be preferred. In this paper, we consider how to infer ensemble properties from the most common quantities available from FRET data, namely the mean ratiometric transfer efficiency[24] and fluorescence lifetimes[25].

As will be discussed in more detail below, we are considering the situation in which the FRET transfer efficiency between donor and acceptor chromophores may be considered to depend only on the distance R between them. In this case, the average transfer efficiency can be obtained by integrating over a given distribution of distances $P(r)$, ideally characterized by a small number of parameters, since often only the mean transfer efficiency is available as an observable. Perhaps the most widely used distribution is a Gaussian (ideal) chain, which has the advantage of having only a single adjustable parameter, usually taken to be the mean-squared distance, $R^2 \equiv \langle r^2 \rangle$, between FRET donor and acceptor. [26] The average radius of gyration, R_g, can in turn be estimated from R using the properties of the polymer model employed. However, it has been recognized that using a Gaussian chain may distort the apparent inter-chromophore distance R when the unfolded protein concerned is close to the excluded volume (EV) limit (scaling exponent ~2/3)[27-29]. In this case, the very broad $P(r)$ for a Gaussian chain contains an unphysical contribution at shorter distances which would tend to increase the apparent mean FRET efficiency (Fig. 1). In order to match the experimental data for a polypeptide with the properties of a pure EV chain (with no attractive intramolecular interactions), an artificially enlarged R^2 would need to be chosen. This artifact is widely acknowledged, and it has been proposed that using a self-avoiding walk (SAW) model is a better description for protein chains at high denaturant concentrations, where they are typically close to the EV limit [28-31]. In a recent study of an unfolded and an intrinsically disordered protein in chemical denaturants, we indeed found that applying a SAW model to FRET data at high denaturant concentrations led to inferred average properties (R and R_g) close to those obtained by a global analysis integrating both FRET and SAXS data in an ensemble description based on atomistic simulations, while the result from using a Gaussian chain overestimated these properties[4]. On the other hand, at low denaturant concentrations, R and R_g inferred from a Gaussian chain model were closer to those from a global analysis than the SAW model. This trend is physically expected, given that unfolded proteins in water often have a scaling exponent characteristic of the Θ-state [31-33](~1/2), where an ideal chain description is expected to work best[31, 34]. This implies that, while using the same polymer model to infer average distances and R_g from FRET data over a range of solvent conditions may seem consistent, doing so may distort the results because the one chosen model is not valid over the complete range of conditions.

For a given protein, it is in principle possible to determine the polymer scaling exponent which best describes its properties under given conditions of solvent and temperature, for example by attaching FRET labels at different positions, separated by a variety of chain lengths[31]. This would allow the appropriate distribution $P(r)$ to be identified. However, the range of accessible sequence separations is limited by the Förster radii of suitable dye pairs, and producing multiple labeling variants presents a considerable additional experimental burden. Ideally, we desire a method that can be used to infer accurate ensemble properties of the chain from a minimum of available experimental data. In this paper, we propose such a method, based on an approximate theoretical distribution parameterized by the scaling exponent ν of the chain. We test the proposed method against synthetic FRET data calculated from known conformational ensembles. These ensembles are generated using an implicit-solvent model including both excluded volume repulsion and as well as realistic attractive interactions within the chain[35], or using an all-atom explicit-solvent model with an optimized force field for intrinsically disordered proteins[36]. We further test the method using experimental FRET data of proteins for which SAXS data are also available[4], so that the reference radius of gyration can be determined more accurately by two different experimental methods. In all cases considered, we find that our method yields a radius of gyration in good agreement with the reference value, while simultaneously giving an accurate estimate of the polymer scaling exponent.

II. Methods

To set the stage, we first provide a basic description of the principles behind a FRET experiment (FIG. 1, top panel). A donor and an acceptor chromophore are covalently linked to the molecule of interest at specific sites. The donor chromophore is optically excited, and the excitation energy is either emitted as a photon, or transferred to an acceptor chromophore, which then emits a photon. Non-radiative processes can also contribute to the donor or acceptor decay, but they are corrected for in the experiments, so we neglect them here. The efficiency of energy transfer is related to both the distance between the chromophores, as well as their relative orientation. We consider only the situation in which the chromophores are rapidly sampling different orientations on the time scale of the average duration of the donor excited state (or donor lifetime)[37], as is frequently the case, especially for unfolded and intrinsically disordered proteins[27]. In this situation, the orientational contribution is averaged out, and the transfer efficiency of a given protein configuration with chromophores separated by distance r is given by:

$$E(r) = \frac{1}{1+(r/R_0)^6}\,, \tag{1}$$

where R_0 is the spectroscopically determined Förster radius[38]. Because the transfer efficiency is an average over a highly heterogeneous conformational ensemble, a commonly used method of analysis averages $E(r)$ over a distribution of distances $P(r;\{\xi_i\})$ computed from a polymer model characterized by a set of parameters $\{\xi_i\}$:

$$\langle E\rangle = \int_0^\infty P(r;\{\xi_i\})E(r)dr \tag{2}$$

A complementary piece of information is the donor fluorescence lifetime[25], i.e., the mean time between excitation of the donor and emission of a donor photon, which can be related to the mean transfer efficiency $\langle E\rangle$ and its variance σ_c^2 by[39, 40]

$$\langle\tau\rangle = \tau_D\left[1-\langle E\rangle + \frac{\sigma_c^2}{1-\langle E\rangle}\right] \tag{3}$$

where the variance of the transfer efficiency, σ_c^2, is:

$$\sigma_c^2 = \int_0^\infty E(r)^2P(r)dr - \langle E\rangle^2 \tag{4}$$

Note that this analysis requires the inter-dye distance relaxation time to be much longer than the donor fluorescence lifetime, which is usually the case for the chromophores used and the chain lengths accessible in single-molecule FRET investigations of unfolded and intrinsically disordered proteins.[27]

In the case of FRET measurements on less heterogeneous systems, such as between chromophores attached to a folded protein, the distance distribution is reasonably narrow and hence sufficiently well described by its mean and variance. However, the end-end distance distribution sampled by an unstructured biopolymer is asymmetric and extremely broad, so that the shape and the tails of the distribution can have a significant effect on the mean transfer efficiency. To illustrate this effect, we show in FIG. 1 example distance distributions for two polymer models, the Gaussian chain and a self-avoiding walk, corresponding to the same mean FRET efficiency. As is evident, the root-mean-square end-end distance $\langle r^2\rangle^{1/2} \equiv R$ of these distributions can differ considerably between the two polymer models when the mean efficiency is identical, particularly at low $\langle E\rangle$, i.e., very expanded chains. Consequently, using a Gaussian chain to interpret a given set of variations in $\langle E \rangle$ would lead to larger inferred variations in R than use of a self-avoiding walk, a fact which is now well appreciated[4, 28, 29]. While the question of which of these (or other) models is a more accurate reflection of the true distribution has received some attention, neither model is likely to provide a fully adequate description of $P(r)$ across all conditions.

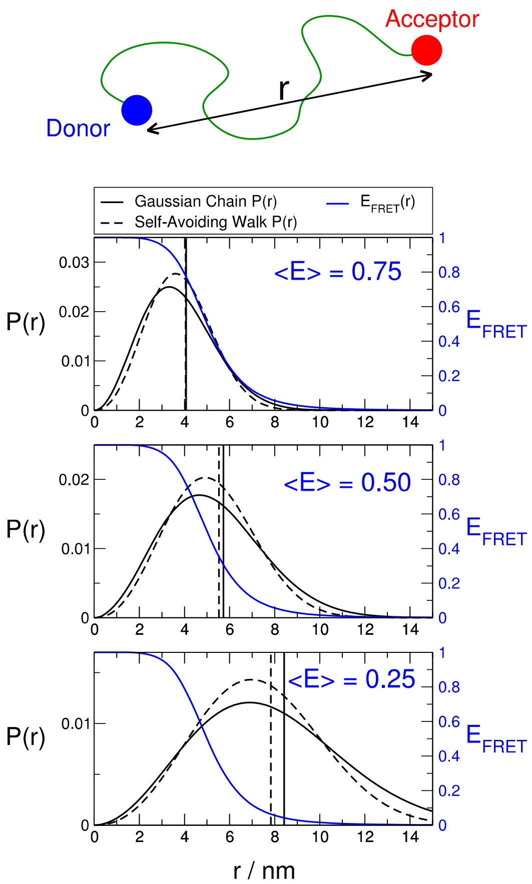

FIG. 1. Determining ensemble properties from mean FRET efficiencies. The top panel illustrates the distance r probed between donor and acceptor attached to a biopolymer of interest. In the lower panels, Gaussian chain and self-avoiding walk end-end distance distributions $P(r)$ are shown for scenarios in which the mean FRET efficiency is 0.75 (top), 0.5 (middle) and 0.25 (bottom). Vertical lines indicate the root mean square end-end distances $R \equiv \langle r^2 \rangle^{1/2}$ for the corresponding $P(r)$. $R_0 = 5$ nm.

In this paper, we propose to use a more general form for the $P(r)$ of a self-avoiding walk (a polymer which cannot cross itself) that can accommodate a variation of the scaling exponent ν:[41, 42]

$$P(r) = A\frac{4\pi}{R}\left(\frac{r}{R}\right)^{2+g}\exp\left[-\alpha\left(\frac{r}{R}\right)^{\delta}\right] . \tag{5}$$

In the above expression, $g = (\gamma - 1)/\nu$ [42] in three dimensions ($\gamma \approx 1.1615$),[43] $\delta = 1/(1-\nu)$,[41] and the constants A and α are determined, for given values of ν and R, from the conditions $\int_0^\infty P(r)dr = 1$ and $\int_0^\infty P(r)r^2 dr = R^2$, as required for normalization, and from the definition of R, respectively. For comparison, the distance distribution for a Gaussian chain is:

$$P(r) = \left(\frac{3}{2\pi}\right)^{3/2} \frac{4\pi}{R} \left(\frac{r}{R}\right)^2 \exp\left[-\frac{3}{2}\left(\frac{r}{R}\right)^2\right] \quad . \tag{6}$$

Note that Eq. 5 does not reduce to Eq. 6, even when $\nu = \frac{1}{2}$, due to the factor $\left(\frac{r}{R}\right)^g$, which attenuates $P(r)$ at short distances, as required for a real chain. A value of $\delta > 2$ reduces the contribution of distances greater than R relative to the Gaussian chain. Strictly speaking, Eq. 5 was derived for a real chain in good solvent, but our analysis below indicates that it provides a useful approximation even outside the good solvent regime.

Eq. 5 has two adjustable parameters, ν and R. If both $\langle E \rangle$ and $\langle \tau \rangle$ are available, it may in principle be possible to determine both parameters directly by fitting the model $P(r)$ in Eq. 5 via Eqs. 2 and 3. In the more common situation that only $\langle E \rangle$ is available, we require some additional information. Here, we exploit the knowledge that the mean end-end distance in unfolded proteins approximately follows the scaling law:

$$R = bN^{\nu} \tag{7}$$

The prefactor b has been estimated for proteins to be approximately 0.55 nm, with negligible effects on the value of v if varied within physically reasonable bounds.[31] With this closure relation, it is possible to solve for both R and v. Note that Eq. 7 can also be applied to other biopolymers besides unfolded proteins by using an appropriately determined prefactor.

The above discussion outlines how the average end-end distance R can be determined, but often one would like an estimate of the radius of gyration R_g, which is the main quantity obtained from scattering experiments. Note that we use the symbol R_g to refer to the root mean square ensemble average over the radii of gyration r_g of individual conformations, i.e. $R_g \equiv \langle r_g^2 \rangle^{1/2}$. For this conversion, we employ the approximate relation [44]:

$$\lambda = \frac{R^2}{R_g^2} = \frac{2(\gamma+2\nu)(\gamma+2\nu+1)}{\gamma(\gamma+1)} \tag{8}$$

Eqs 5 and 8 for self-avoiding walks are approximations, and strictly speaking, only applicable to homopolymers with large chain length. Therefore, the accuracy of the extracted parameters needs to be tested for a model which is a closer approximation to a real heteropolymeric protein chain (or another biopolymer). For this purpose, we utilize both molecular simulations and experimental data.

III. Results and Discussion

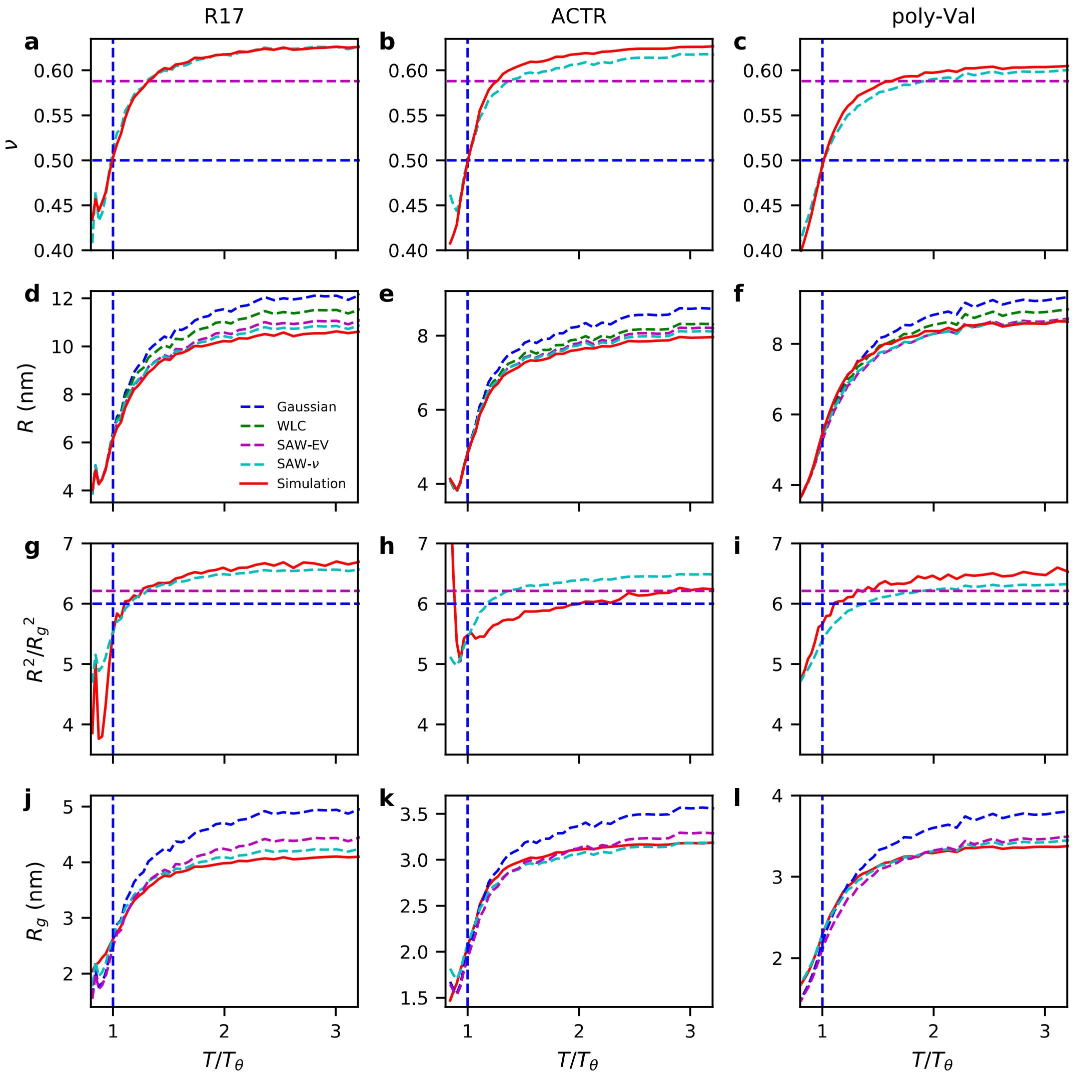

FIG. 2. Recovering ensemble properties from synthetic FRET data calculated from implicit solvent simulations. The recovered properties, obtained by fitting the FRET efficiency $\langle E \rangle$ using different polymer models, are compared with those of the original ensemble used to generate the data, for unfolded R17 (left), ACTR (middle) and poly-Val (right). In (a), (b) and (c) we show the fitted scaling exponent ν, in (d), (e) and (f) the end-end distance R, in (g), (h) and (i) the ratio $\lambda = R^2/R_g^2$ derived from ν, and in (j), (k) and (l) the resulting R_g obtained from R and ν. The legend shows the polymer model used. For the purposes of this plot, T_θ is defined as the temperature for which $\nu = 1/2$. The exponent ν in the simulations is determined from internal distance scaling.[45]

Testing the model with implicit solvent simulation data

First, we use implicit-solvent molecular simulation data from three protein sequences to test the model, including the 116-residue mutant of R17 spectrin domain[4, 46] which stays unfolded even at low denaturant concentration, the 79-residue IDP ACTR[4, 47] (see details in Supporting Methods), and a 100-residue poly-Val peptide. The last sequence is included as a reference homopolymer. While an implicit solvent model is clearly an approximation, we note that the force field used

(ABSINTH[35]) has been shown to capture sequence-dependent variations of the radius of gyration (and hence, of the scaling exponent) reasonably well in previous studies[48, 49]. Dyes are not present in the simulations in the current work as the effect of dyes on the protein dimensions has been discussed in previous publications [50-52]. The positions for end-end distances are thus assumed to be the C^{α} carbon atoms of the terminal residues. By running simulations at different temperatures, we obtained a set of ensembles with different polymer scaling exponents ν covering a wide range from ~0.40 to ~0.65 (FIG. 2a, b and c); note that due to the finite length of the chain, the scaling exponent can exceed the theoretical EV limit of 0.588. The observable $\langle E \rangle$ can be back-calculated from the average of the individual frames of each ensemble (a Förster radius of 5.4 nm is assumed in all cases both for generating synthetic $\langle E \rangle$ and for inferring ensemble properties from it). We utilize $P(r)$ from different polymer models to obtain ν, R and R_g from this value of $\langle E \rangle$ and then compare them with those determined directly from the simulation coordinates. The polymer models tested are the Gaussian chain, the self-avoiding walk in the EV limit (SAW-EV), the worm-like chain (WLC, see Supporting methods for details), and the SAW with variable ν (SAW-ν, Eq. 6 and 7).

In FIG. 2d, e and f, we first show that the distance is similarly well recovered for all the models when ν is close to 0.5; however, for more expanded chains this is no longer the case and a deviation from the mean distance directly calculated from the simulation is observed. In particular, the Gaussian chain model overestimates the distance most, followed by the WLC and SAW-EV. SAW-ν still overestimates the distance by about 2% in both R17 and ACTR for $T/T_{\theta} = 3$, but provides the most accurate values. The increased accuracy of the variable-ν model is due to a more realistic $P(R)$, especially when the chain is more expanded than the excluded volume limit for long chains ($\nu > 0.588$), as shown in FIG. 3 and FIG. S1. These results suggest that the SAW-ν model provides a good approximation not only for homopolymers but also for finite-length heteropolymers, and that a well-chosen $P(R)$ is important for accurately determining chain dimensions from the FRET efficiency.

Importantly, SAW-ν gives the scaling exponent ν in addition to the distance in the procedure of fitting the FRET efficiency. This distinguishes it from the other polymer models used with a constant ν. In FIG. 2a, b and c, we show that the polymer scaling exponent recovered from SAW-ν is in remarkable agreement with that calculated directly from the simulations (see Supporting Methods for details of calculating ν). This observation suggests that the SAW-ν model is self-consistent in obtaining both R and ν. The variable ν, instead of a constant ν as in the Gaussian chain and SAW-EV, is necessary to describe the scaling behavior of real proteins in different solvent conditions.

To calculate R_g from R, one additional parameter is required, which is the ratio λ between R^2 and R_g^2. This parameter is also dependent on the polymer model and has recently been suggested to vary significantly for unfolded proteins in different solvent conditions[53]. In FIG. 2g, h and i, we show that the change in λ computed directly from the simulation is reproduced much better by the SAW-ν model (Eq. 8) than by the constant ratios of 6 and 6.26,[54] respectively, for the Gaussian chain and excluded-volume chain models. As estimated from the SAW-ν model, using a constant λ will introduce an error in R_g of 5% for ν=0.5 and 2% for ν=0.588. The more accurate estimate of λ from the variable- ν model consequently leads to final estimates of R_g (FIG. 2j, k and l) that are in much better agreement with those calculated from the simulations than the two most commonly used polymer models, the Gaussian chain model and SAW-EV. Thus, improvements in the ν-dependent estimates of both $P(R)$ and λ help to infer a more accurate radius of gyration from the FRET efficiency.

While $\nu \in [0.5, 0.588]$ may seem to cover the likely situations for unfolded and disordered proteins, in the case of short chains, ν can lie outside this range, as has been found for real proteins[31]. To our knowledge, a rigorous theoretical description covering the entire range of relevant solvent conditions for the finite-length chains of interest here is thus currently not available. One polymer model that can describe a $P(\mathrm{r})$ varying between theta solvent and excluded volume limits is that of Oono and Freed[55], but the scaling exponent is only allowed to vary between 0.5 and 0.588, so it is difficult to apply it to finite-length chains. We therefore employ Eq. 5 as an approximation for analyzing single-molecule FRET data. Even though the equation is derived to describe a homopolymer, we do not see a big difference in accuracy when applying the SAW-ν model to natural protein sequences (i.e. ACTR and R17) versus a homopolymer sequence (i.e. poly-Val). This observation suggests that unfolded/intrinsically disordered proteins with well-mixed sequences may still be described sufficiently well with models developed for homopolymers.

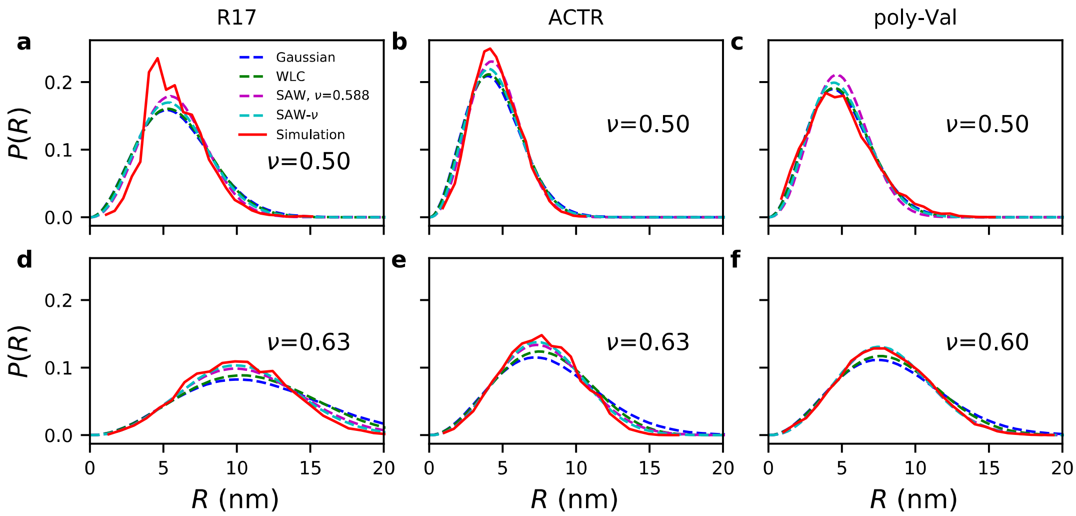

FIG. 3. Examples of distance distributions from different polymer models (see legend) and from the implicit solvent simulations (red) for scaling exponents of ~0.5 (top row) or ~0.6 (bottom row).

Since fluorescence lifetimes also report on the distance distribution $P(r)$ (Eq. 3), they can be used as an additional or alternative observable to determine R and R_g. As for the FRET efficiency, we back-calculate the donor lifetime $<\tau>$ from the simulations, obtain ν, R and R_g using SAW-ν, and then compare them with their counterparts computed directly from the simulations (FIG. S2). The parameters obtained from $<\tau>$ are very similar to those from the ratiometric FRET efficiency. This is not entirely surprising since the variance of the distance distribution is highly correlated with the average distance in the polymer model. The good agreement we observe for the parameters obtained independently from $<\tau>$ and $<E>$ reflects the good agreement between the shapes of $P(r)$ from the SAW-ν model and from the simulations (FIG. 3 and FIG. S1). Lifetime information can be very useful for identifying the presence of a broad distribution of rapidly interconverting distances in the sample in a model-free manner[40, 56] and provides a valuable experimental control[27, 57], but the variances of the different polymer distance distributions are too similar to be discriminated reliably (FIG. 3). As a result, a combined analysis of $<\tau>$ and $<E>$ does not allow both the scaling prefactor b and the scaling exponent ν to be determined independently. Here we thus limit our analysis to the mean transfer efficiencies.

Since one experimental observable, the mean transfer efficiency, only allows us to determine one free parameter of the model, we have used a fixed scaling prefactor b taken from earlier FRET experiments[31]. While one might try to determine b by fitting the average distance R between residue pairs i,j as a function of their sequence separation $|i-j|$ (FIG. S3) to $R = b|i-j|^{\nu}$, correlations between b and ν make it difficult to accurately determine both parameters independently, especially for the limited range of sequence separations typically accessible for a single protein (a complication that is valid both for the simulations and, even more so, for the experimental data). A prefactor of b = 0.55 is suitable for recovering the properties of the original ensemble when used in our SAW-ν model. We show in FIG. S4 and S5 the results for prefactors of 0.45 and 0.65 nm, respectively, to determine ensemble properties via the SAW-ν model. While the recovered end-end distances are quite robust and thus similar to those obtained using a prefactor of 0.55 nm (FIG. 2), a smaller prefactor tends to overestimate the ratio λ between R^2 and R_g^2, whereas a larger prefactor underestimates λ, leading to corresponding errors in R_g. This observation suggests that a prefactor of 0.55 nm is close to optimal for use in the SAW-ν model.

Testing the model with all-atom explicit solvent simulation data

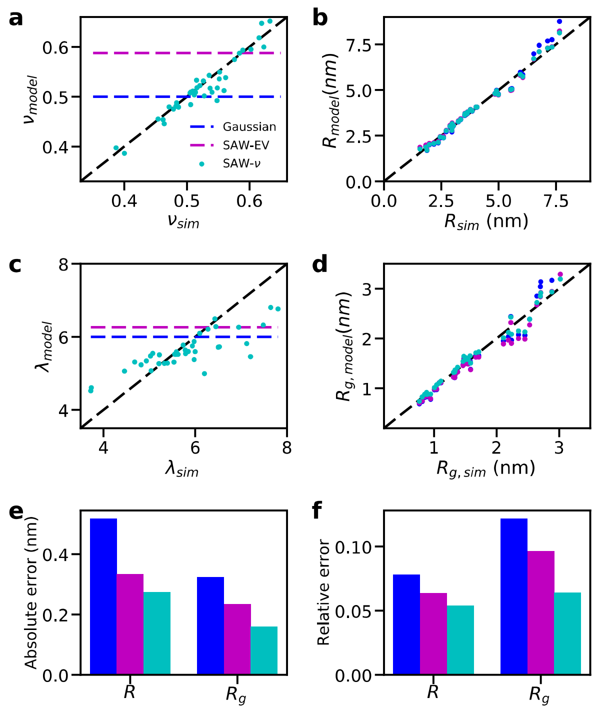

FIG. 4. Recovering ensemble properties from synthetic FRET data, with $\langle E \rangle$ calculated from explicit solvent simulations. We plot the recovered properties versus those of the original ensemble for (a) the scaling exponent ν, (b) the root-mean-squared end-end distance, R, (c) $\lambda = R^2/R_g^2$ and (d) the radius of gyration, R_g. Errors in (e) and (f) are computed as root mean square averages over the different proteins. The absolute error defined as $\langle (x_{\mathrm{model}} - x_{\mathrm{sim}})^2 \rangle^{1/2}$ is shown in (e) and the relative error $\langle ((x_{\mathrm{model}} - x_{\mathrm{sim}})/x_{\mathrm{sim}})^2 \rangle^{1/2}$ in (f). The legend shows the polymer model used, i.e. Gaussian chain model (blue), SAW-EV (magenta) and SAW-ν (cyan).

Our tests so far are based on simulations of unstructured proteins in implicit solvent. However, simulation models in which the solvent is represented explicitly are in principle the most realistic. We have taken advantage of a large recently obtained set of explicit solvent trajectories using the force field Amber03ws[36] for proteins, TIP4P/2005[58] for water and (in some cases) KBFFs[59] for denaturants. Previous simulations with the Amber03ws model have resulted in good agreement with FRET[36, 60], SAXS[36, 60, 61] and contact quenching[62] experiments for diverse protein sequences. A particular strength is that it captures local structure formation well[36, 63], in addition to global properties such as radius of gyration. While there are some deficiencies with respect to the secondary structure propensity of individual amino acids[64, 65], for the present purposes these would only matter for sequences with a preponderance of those residues, i.e. not for the well-mixed sequences studied here. We used 39 trajectories in total from 30 proteins covering a chain length from 11 to 89 in different solvent conditions (i.e. water, urea and GdmCl) (Table S1). These trajectories cover a wide range of scaling exponents from 0.39 to 0.63, which were determined by fitting the dependence of pairwise distances between C^{α} atoms on sequence separation (FIG. S3). A full list of the simulations and proteins used is given in Table S1, with details in Supporting methods. These force fields have been specifically developed with the aim of accurately representing ensembles of unfolded and disordered proteins, and have been validated in several independent studies[36, 59, 63, 66, 67]. We therefore believe that these simulations generate more realistic conformational ensembles of unfolded/intrinsically disordered proteins than most other explicit solvent force fields, which typically yield structures that are too collapsed[68]. The simulation data have been collected either with unbiased MD simulations of length much greater than typical relaxation times for global parameters such as R_{g}, or via enhanced sampling methods, as detailed in the SI Text, in order to obtain representative samples of the equilibrium distributions. For each trajectory, the same framework used in testing the implicit solvent ensembles is applied.

In FIG. 4, we show the accuracy of recovering the scaling exponent, end-end distance and radius of gyration using synthetic FRET efficiencies calculated from the end-end distances in the original trajectories, compared with the parameters directly obtained from the trajectories. All three polymer models (Gaussian chain, SAW-EV and SAW-ν) show a good correlation between the recovered and directly calculated parameters (FIG. 4b and d). However, only SAW-ν is able to capture the changing scaling exponent and ratio λ between R^2 and R_g^2, and a comparison of the absolute errors of the three models shows that SAW-ν always performs best (FIG. 4e). We have found no correlation between the chain length and the relative errors (FIG. S6), which suggests that SAW-ν, which was derived for long homopolymers, is not sensitive to the chain length once it is greater than ~10. However, the model starts to break down when the chain length is less than ~10, as reflected by a poor fit of intrachain distances as a function of sequence separation to Flory's scaling law (FIG. S7).

Testing the model with experimental data

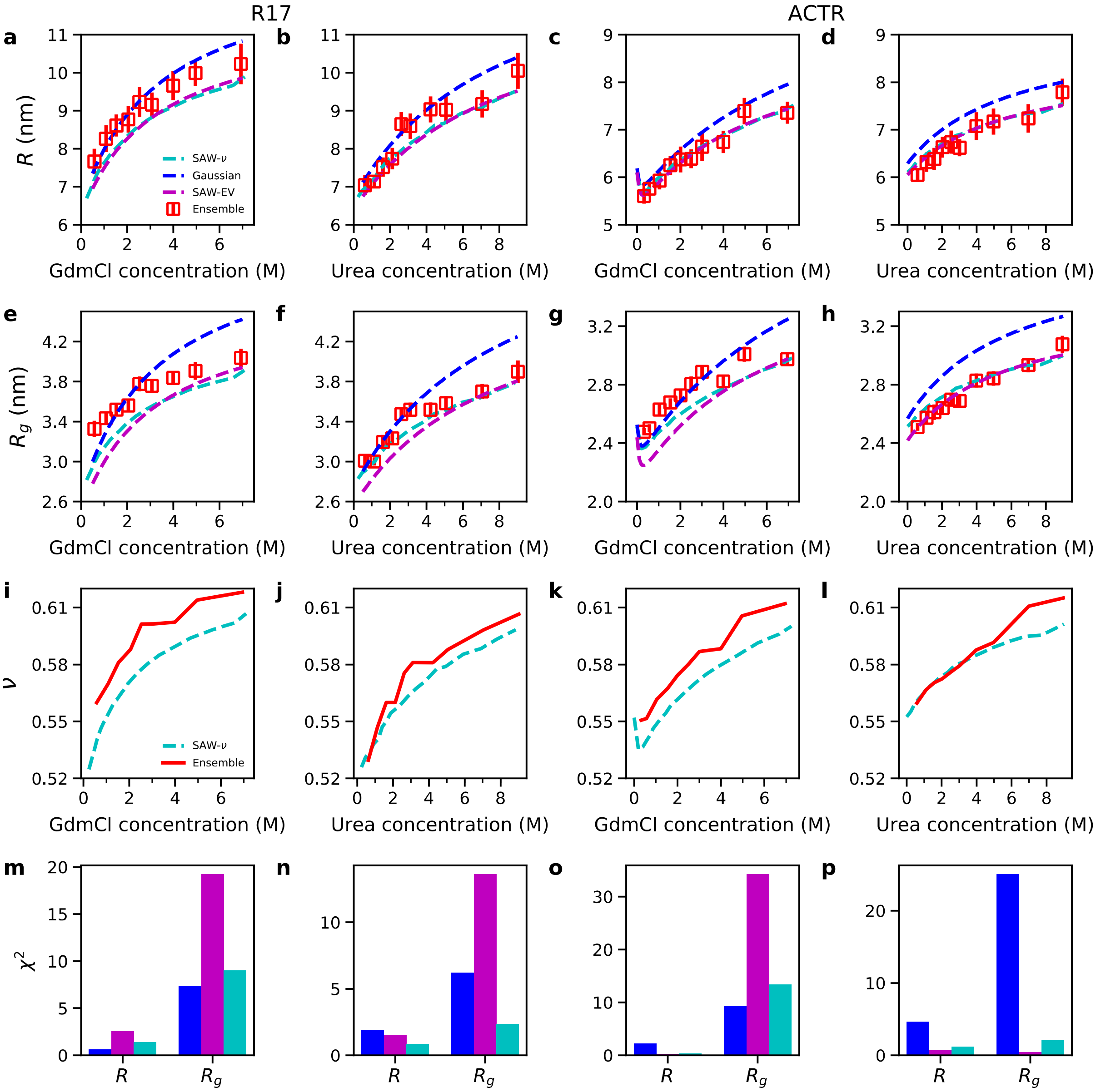

FIG. 5. Validation using ensembles determined from both SAXS and FRET data (proteins labeled with Alexa 488 and 594, R_0 = 5.4 nm) at different denaturant concentrations (GdmCl, urea). Ensemble properties estimated using polymer models from the FRET data alone are compared with those computed from molecular ensembles ("Ensemble" in legend) generated to match both SAXS and FRET data. We show R, R_g, and the scaling exponent recovered from FRET efficiency at each denaturant concentration, and χ^2 defined as $<((x_{\text{ensemble}} - x_{\text{sim}})/\sigma_{\text{ensemble}})^2>$ of all denaturant concentrations for the proteins R17 and ACTR. The legend shows the polymer model used, i.e. Gaussian chain model is in blue; SAW-EV in magenta and SAW-ν in cyan.

While we consider the simulation ensembles we have studied to be a realistic representation of unfolded states under various conditions, they are nonetheless based entirely on an empirical energy function. One would ideally like to use experimental ensembles, but determining such ensembles from experimental data alone is currently not possible. Instead, experimental data can be used in conjunction with a reasonably accurate force field to generate an ensemble, with the experimental data helping to correct deficiencies in the force field. In a recent publication[4], we have performed both SAXS and FRET measurements on the unfolded R17 and the IDP ACTR. We used an ensemble reweighting method[4, 12] to obtain a unified molecular description of both experimental data sets, together with a simulation model (the ABSINTH force field[35]). In these ensembles, the radius of gyration and end-end distance are strongly constrained by the SAXS and FRET data used to generate them, and rely less on the properties of the force field. The reweighted ensembles from both SAXS and FRET experimental data therefore serve as a reference that can be used to test the accuracy of the SAW-ν model in interpreting real experimental data. This comparison can also tell us whether a good estimate of R_g could have been obtained by a simpler method than ensemble reweighting.

In FIG. 5, we present the comparison with the reweighted ensembles. Note that we have four data sets comprising unfolded R17 in urea, unfolded R17 in guanidinium chloride (GdmCl), ACTR in urea, and ACTR in GdmCl. On average, the SAW-ν model provides the most accurate estimate of R_g and R over all experimental conditions (Table S2). The SAW-EV model performs much worse than the other two models in three cases, whereas the Gaussian chain model performs much worse in one case; this is likely to be a consequence of the respective systems being closer to the ideal- or EV-chain limits. By including the variability in the scaling exponent explicitly in the model, SAW-ν is better able to describe all of these cases, especially the ones exhibiting a crossover from near-Θ to good solvent conditions upon increasing denaturant concentration. The most significant discrepancy is a slight underestimation of the scaling exponent for the experiments in guanidinium chloride. This might be explained by a small change of prefactor due to the addition of the cosolvent, an effect which might be included via cosolvent-dependent prefactor in future versions of the model. We have also compared the model with reference ensembles obtained only from FRET measurements, since the method is introduced here primarily for the analysis of FRET experiments. In FIG. S8, we show that the SAW-ν yields a χ^2 less than 1 in most cases, suggesting that the method is appropriate to achieve an estimate of R and R_g within experimental error. In addition to R_g, the scaling exponent obtained from SAW-ν, which is a constant in the other models, is in reasonable agreement with that calculated directly from the reweighted ensemble. This suggests that SAW-ν captures the scaling behavior of the protein, and that a ν-dependent $P(R)$ provides an improved description of unfolded or intrinsically disordered proteins compared to models assuming ν to be constant.

IV. Conclusion

We have demonstrated a straightforward scheme to estimate more accurately the end-end distance R and radius of gyration R_g directly from FRET measurements on unfolded proteins. Not only is the scheme more accurate than polymer models commonly used to date, but it also returns the scaling exponent of the chain. These advantages come partly via the assumption of a given scaling prefactor. The value we have chosen is based on experimental estimates, and its use has been validated here via the agreement between inferred properties such as R, R_g, R^2/R_g^2 and ν and their true values in our simulation ensembles. Thus it appears that the variation of the optimal prefactor for typical IDP sequences is small; a possible direction for future improvement may be the inclusion of a sequence-dependent prefactor to accommodate unusual sequences and the effects of different co-solvents. For example, we find that a smaller prefactor of 0.45 nm is better suited for extracting properties of a poly-Ala sequence via SAW-ν (results not shown). Nonetheless, it appears that a common value of 0.55 nm is a good first approximation for typical IDP sequence compositions, as the ones studied here. Similarly, in order to apply the method to other biomolecules, for example RNA, it would be necessary to determine the prefactor for those cases. This could be done either experimentally[31] or via simulation models such as those used here. In addition, our approach can also be applied to interpreting intramolecular distances from other experimental methods, if the system of interest can be described by the polymer model.

Like the Gaussian chain and SAW-EV models, the SAW-ν model is strictly applicable only to homopolymers. Proteins are heteropolymers, and it has been shown that the patterning of sequence features such as charge can have significant effects on their global properties[69], which clearly would not be captured by any of the homopolymer theories discussed here. Heteropolymer effects have also been suggested to be important for explaining the discrepancy between SAXS and FRET experiments as applied to denatured proteins[53, 70, 71]. Incorporating such effects would in general require a hybrid approach in which the experimental data are used to refine a reasonably accurate simulation ensemble. However, running such simulations adds computational cost and complexity to the analysis, making it less widely accessible. Provided that the protein under consideration has a reasonably well-mixed sequence, it should still be possible to approximate its global properties with a homopolymer theory. How well-mixed does the sequence need to be? We have studied here a wide range of intrinsically disordered and unfolded proteins, and we find empirically that indeed the approximate homopolymer theory applied here is sufficient to obtain remarkably accurate estimates in all cases considered. In fact, application of the SAW-ν model to a poly-Val homopolymer yields results of similar accuracy for R and R_g as those obtained for the natural protein sequences. Of course, a pronounced patterning of sequence properties as found in some IDPs[72] likely would cause the method to fail. Signs of such cases could be a fitted scaling exponent outside the range [0.4-0.65] most commonly observed for unfolded and disordered proteins, or a discrepancy between properties estimated from the ratiometric FRET efficiency and the donor fluorescence lifetime. The most stringent experimental test is to probe different segments of the chain by varying the positions of FRET donor and acceptor in the protein and seeing whether they can be described globally with a single set of fit parameters[4, 28, 73, 74]. Discrepancies would suggest that a homopolymer $P(r)$ cannot provide a self-consistent interpretation of the data and that more detailed analysis is needed, including, e.g., atomistic simulations or ensemble reweighting, ideally combined with additional segment-specific experimental information, e.g. from NMR[3, 74].

The SAW-ν method allows a scaling exponent to be determined using a FRET measurement from a single labeled variant of a protein, in addition to a more accurate estimate of the dye separation R and the radius of gyration R_g than established methods. This raises the question of which is the optimal pair of residues to label, from the perspective of this method? We suggest based on our results that choosing a pair such that their average separation R is comparable to the Förster radius R_0 yields the most accurate ensemble properties. Choosing an average separation much larger or smaller than R_0 will make the results more sensitive to the accuracy of the tails of the distance distribution $P(r)$, and hence less robust than choosing R close to R_0.

After submission of our work, a novel method of analyzing small-angle X-ray scattering (SAXS) experiments was published which also allows a scaling exponent to be determined from a single experiment[75]. An approach similar to the SAW-ν model should also be applicable to analyzing SAXS experiments to obtain both the scaling exponent and R_g. We are currently investigating this possibility.

V: SUPPLEMENTARY MATERIAL

See supplementary material for supporting methods, figures and tables.

ACKNOWLEDGMENTS

We thank Andrea Soranno for many insightful discussions. R.B. and W.Z. were supported by the intramural research program of the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). B.S. was supported by the Swiss National Science Foundation.

REFERENCES

1. P. E. Wright and H. J. Dyson, Nat. Rev. Mol. Cell Biol. 16, 18-29 (2015).
2. D. H. Brookes and T. Head-Gordon, J Am Chem Soc 138 (13), 4530-4538 (2016).

3. M. Schwalbe, V. Ozenne, S. Bibow, M. Jaremko, M. Gajda, M. Ringkjøbing-Jensen, J. Biernat, S. Becker, E. Mandelkow, M. Zweckstetter and M. Blackledge, Structure 22, 238-249 (2014).
4. A. Borgia, W. Zheng, K. Buholzer, M. B. Borgia, A. Schuler, H. Hofmann, A. Soranno, D. Nettels, K. Gast, A. Grishaev, R. B. Best and B. Schuler, J Am Chem Soc 138 (36), 11714-11726 (2016).
5. S. Meier, M. Blackledge and S. Grzesiek, J. Chem. Phys. 128, 052204 (2008).
6. C. C. Jao, A. Der-Sarkissian, J. Chen and R. Langen, Proc Natl Acad Sci U S A 101 (22), 8331-8336 (2004).
7. A. Nöppert, K. Gast, M. Müller-Frohne, D. Zirwer and G. Damaschun, FEBS Lett. 380, 179-182 (1996).
8. T. Dertinger, V. Pacheco, I. Von der Hocht, R. Hartmann, I. Gregor and J. Enderlein, Chem. Phys. Chem. 8, 433-443 (2007).
9. M. A. Graewert and D. I. Svergun, Curr. Opin. Struct. Biol. 23, 748-754 (2013).
10. X. Michalet, S. Weiss and M. Jager, Chem Rev 106 (5), 1785-1813 (2006).
11. B. Schuler and H. Hofmann, Curr Opin Struct Biol 23 (1), 36-47 (2013).
12. G. Hummer and J. Koefinger, J Chem Phys 143 (24) (2015).
13. O. F. Lange, N.-A. Lakomek, C. Fares, G. F. Schröder, K. F. A. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger and B. L. De Groot, Science 320, 1471-1475 (2008).
14. E. Boura, B. Rózycki, D. Z. Herrick, H. S. Chung, J. Vecer, W. A. Eaton, D. S. Cafiso, G. Hummer and J. H. Hurley, Proc. Natl. Acad. Sci. U. S. A. 108, 9437-9442 (2011).
15. J. R. Allison, R. C. Rivers, J. C. Christodoulou, M. Vendruscolo and C. M. Dobson, Biochemistry 53, 7170-7183 (2014).
16. C. K. Fisher, A. Huang and C. M. Stultz, J. Am. Chem. Soc. 139, 14919-14927 (2010).
17. W. Boomsma, J. Ferkinghoff-Borg and K. Lindorff-Larsen, PLoS Comput. Biol. 10, e1003406 (2014).
18. K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson and M. Vendruscolo, Nature 433, 128-132 (2005).
19. R. B. Best and M. Vendruscolo, in *J. Am. Chem. Soc.* (2004), Vol. 126, pp. 8090-8091.
20. J. Kuriyan, K. Osapay, S. K. Burley, A. T. Brunger, W. A. Hendrickson and M. Karplus, in *Proteins* (1991), Vol. 10, pp. 340-358.
21. V. Venditti, T. Egner and G. M. Clore, Chem. Rev. 116 (11), 6305-6322 (2016).
22. W. Rieping, M. Habeck and M. Nilges, in *Science* (2005), Vol. 309, pp. 303-306.
23. B. Roux and J. Weare, J. Chem. Phys. 138 (084107) (2013).
24. A. A. Deniz, T. A. Laurence, M. Dahan, D. S. Chemla, P. G. Schultz and S. Weiss, Annu. Rev. Phys. Chem. 52, 233-253 (2001).
25. E. Sisamakis, A. Valeri, S. Kalinin, P. J. Rothwell and C. A. M. Seidel, Methods Enzymol 475, 455-514 (2010).
26. P. J. Flory, *Principles of Polymer Chemistry*. (Cornell University Press, Ithaca and London, 1953).
27. B. Schuler, A. Soranno, H. Hofmann and D. Nettels, Annu Rev Biophys 45, 207-231 (2016).
28. E. O'Brien, G. Morrison, B. R. Brooks and D. Thirumalai, J. Chem. Phys. 130, 124903 (2009).
29. J. Song, G.-N. Gomes, C. C. Gradinaru and H.-S. Chan, J. Phys. Chem. B 119, 15191-15202 (2015).
30. J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach and a. K. W. Plaxco, in *Proc. Natl. Acad. Sci. U.S.A.* (2004), Vol. 101, pp. 12491-12496.
31. H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels and B. Schuler, Proc Natl Acad Sci U S A 109 (40), 16155-16160 (2012).
32. B.-Y. Ha and D. Thirumalai, Phys. Rev. A 46 (6), R3012-R3015 (1992).
33. A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine and R. V. Pappu, Proc. Natl. Acad. Sci. U. S. A. 107, 8183-8188 (2010).

34. K. A. Merchant, R. B. Best, J. M. Louis, I. V. Gopich and W. A. Eaton, Proc. Natl. Acad. Sci. U.S.A. 104, 1528-1533 (2007).
35. A. Vitalis and R. V. Pappu, J. Comput. Chem. 30, 673-699 (2008).
36. R. B. Best, W. Zheng and J. Mittal, J. Chem. Theor. Comput. 10, 5113-5124 (2014).
37. B. Schuler, E. A. Lipman, P. J. Steinbach, M. Kumke and W. A. Eaton, Proc. Natl. Acad. Sci. U. S. A. 102, 2754-2759 (2005).
38. T. Förster, Ann. Phys. 6, 55-75 (1948).
39. I. V. Gopich and A. Szabo, Proc. Natl. Acad. Sci. U. S. A. 109, 7747-7752 (2012).
40. H. S. Chung, J. M. Louis and I. V. Gopich, J. Phys. Chem. B 120, 680-699 (2016).
41. M. E. Fisher, J. Chem. Phys. 44, 616-622 (1966).
42. J. des Cloizeaux, Phys Rev A 10 (5), 1665-1669 (1974).
43. J. C. Le Guillou and J. Zinn-Justin, Phys Rev Lett 39 (2), 95-98 (1977).
44. T. A. Witten and L. Schäfer, J. Phys. A 11 (9), 1843-1854 (1978).
45. H. T. Tran, A. Mao and R. V. Pappu, J. Am. Chem. Soc. 130, 7380-7392 (2008).
46. A. Borgia, B. G. Wensley, A. Soranno, D. Nettels, M. B. Borgia, A. Hoffmann, S. H. Pfeil, E. A. Lipman, J. Clarke and B. Schuler, Nat Comm 3, 1195 (2012).
47. S. J. Demarest, M. Martinez-Yamout, J. Chung, H. Chen, W. Xu, H. J. Dyson, R. M. Evans and P. E. Wright, Nature 415, 549-553 (2002).
48. R. Wuttke, H. Hofmann, D. Nettels, M. B. Borgia, J. Mittal, R. B. Best and B. Schuler, Proc. Natl. Acad. Sci. U. S. A. 111, 5213-5218 (2014).
49. R. K. Das and R. V. Pappu, Proc. Natl. Acad. Sci. U. S. A. 110, 13392-13397 (2013).
50. R. B. Best, H. Hofmann, D. Nettels and B. Schuler, Biophys J 108 (11), 2721-2731 (2015).
51. W. Zheng, A. Borgia, K. Buholzer, A. Grishaev, B. Schuler and R. B. Best, J Am Chem Soc 138 (36), 11702-11713 (2016).
52. G. H. Zerze, R. B. Best and J. Mittal, Biophys. J 107, 1654-1660 (2014).
53. G. Fuertes, N. Banterlea, K. M. Ruff, A. Chowdhury, D. Mercadante, C. Koehler, M. Kachala, G. E. Girona, S. Milles, A. Mishra, P. R. Onck, F. Grater, S. Esteban-Martin, R. V. Pappu, D. I. Svergun and E. A. Lemke, Proc Natl Acad Sci U S A 114 (31), E6342-E6351 (2017).
54. L. Schäfer, *Excluded volume effects in polymer solutions: as explained by the renormalization group*. (Springer Science & Business Media, 2012).
55. Y. Oono and K. F. Freed, J Phys a-Math Gen 15 (6), 1931-1950 (1982).
56. S. Kalinin, A. Valeri, M. Antonik, S. Felekyan and C. A. M. Seidel, J Phys Chem B 114 (23), 7983-7995 (2010).
57. A. Soranno, B. Buchli, D. Nettels, R. R. Cheng, S. Muller-Spath, S. H. Pfeil, A. Hoffmann, E. A. Lipman, D. E. Makarov and B. Schuler, Proc Natl Acad Sci U S A 109 (44), 17800-17806 (2012).
58. J. L. F. Abascal and C. Vega, J. Chem. Phys 123, 234505 (2005).
59. W. Zheng, A. Borgia, M. B. Borgia, B. Schuler and R. B. Best, J. Chem. Theor. Comput. 11, 5543-5553 (2015).
60. W. Zheng, A. Borgia, K. Buholzer, A. Grishaev, B. Schuler and R. B. Best, J. Am. Chem. Soc. 138, 11702–11713 (2016).
61. J. Henriques, C. Cragnell and M. Skepö, J. Chem. Theor. Comput. 11, 3420-3431 (2015).
62. G. H. Zerze, J. Mittal and R. B. Best, Phys. Rev. Lett. 116, 068102 (2016).
63. A. E. Conicella, G. H. Zerze, J. Mittal and N. L. Fawzi, Structure 24, 1537-1549 (2016).
64. R. B. Best, D. De Sancho and J. Mittal, Biophys. J 102, 1462-1467 (2012).
65. F. Meng, M. M. J. Bellaiche, J.-Y. K. Kim, G. H. Zerze, R. B. Best and H. S. Chung, Biophys. J., in press (2017).
66. J. Henriques, C. Cragnell and M. Skepo, J Chem Theory Comput 11 (7), 3420-3431 (2015).
67. G. H. Zerze, J. Mittal and R. B. Best, Phys Rev Lett 116 (6), 068102 (2016).

68. S. Piana, J. L. Klepeis and D. E. Shaw, Curr. Opin. Struct. Biol. 24, 98-105 (2014).
69. R. K. Das and R. V. Pappu, Proc Natl Acad Sci U S A 110 (33), 13392-13397 (2013).
70. K. M. Ruff and A. S. Holehouse, Biophys J 113 (5), 971-973 (2017).
71. J. Song, G. N. Gomes, T. Shi, C. C. Gradinaru and H. S. Chan, Biophys J 113 (5), 1012-1024 (2017).
72. R. K. Das, K. M. Ruff and R. V. Pappu, Curr Opin Struct Biol 32, 102-112 (2015).
73. A. Hoffmann, A. Kane, D. Nettels, D. E. Hertzog, P. Baumgartel, J. Lengefeld, G. Reichardt, D. A. Horsley, R. Seckler, O. Bakajin and B. Schuler, Proc Natl Acad Sci U S A 104 (1), 105-110 (2007).
74. M. Aznauryan, L. Delgado, A. Soranno, D. Nettels, J. R. Huang, A. M. Labhardt, S. Grzesiek and B. Schuler, Proc Natl Acad Sci U S A 113 (37), E5389-E5398 (2016).
75. J. A. Riback, M. A. Bowman, A. M. Zmyslowski, C. R. Knoverek, J. M. Jumper, J. R. Hinshaw, E. B. Kaye, K. F. Freed, P. L. Clark and T. R. Sosnick, Science 358 (6360), 238-241 (2017).